§20. On the Construction of Databases of Experiment Data

Hochin, T. (Osaka Prefecture Univ.)
Nakanishi, H., Kojima, M.

Experiments of the fusion phenomena produce a lot of sequences of time-varying values. A sequence of the values forms a waveform. If the waveforms similar to a desired one can be obtained by using computer system, the burden of researchers in searching similar waveforms will be extremely decreased. Finding the similar waveforms may bring us new breakthrough. We have addressed to the issue on this kind of retrieval [1]. The proposed method is based on the Fourier Transformation [1]. The first several Fourier coefficients are used to narrow the search space. The candidates obtained are again evaluated by using high dimensions' and distinctive coefficients. This method is called the *two-step method*. This method is designed for the whole matching of waveforms.

Another type of matching of waveforms is the subsequence matching. In the subsequence matching, the query sequence is smaller than the evaluated sequences. A sequence is searched in the large sequence that best matches the query sequence. The subsequence matching of waveforms is also strongly required. This paper addresses to the method of the subsequence matching.

The simplest way of the subsequence matching of waveforms is that every point of a waveform is compared with that of a query waveform with shifting the retrieval start position to the next point. This method is simple, but retrieval performance will not be tolerable. The method with good retrieval performance is required.

One of the way of improvement is that a waveform is divided into segments, and the similarity of a waveform is evaluated by using those of the component segments. There may be several methods of evaluating the similarity. We use the two-step method in evaluating the similarity of each segment. This method is called the *simple two-step method*. From here on, the first segment is first evaluated, and the following segments are evaluated according to their order in a waveform for the simplicity.

Following the simple two-step method, the more the number of segments composing a query waveform is, the less the number of the candidates obtained is. No candidates can be obtained when a query waveform is long. The method is improved to keep the number of the candidates. That is, when the number of candidates is smaller than the pre-defined lower bound, the method asks the multi-dimensional index to return more candidates and set the lower bound to the smaller one in evaluating each segment. This method is called the *candidate-keep two-step method*.

The methods described above use multi-dimensional index in evaluating every segment. When one segment that may be the component of a required waveform can be obtained, similarity of the waveform can be evaluated by consulting the segment next and/or prior to that segment. There are two approaches to this evaluation. One approach modifies the structure of the multi-dimensional index to be able to reach directly to the next or prior segment. The other keeps the information on the neighbor segments out side of the multi-dimensional index. The next or prior element is evaluated by using the information out side of the multi-dimensional index. We adopt the latter approach because we do not have to change the multi-dimensional index. Fourier coefficients of the segments that are the same ones as those used in the two-step method are stored into a file, which is called a binary file, out side of the multi-dimensional index. The first segment is evaluated by using the multi-dimensional index. The following segments are evaluated by using the binary file. This method is called the *binary file metho*d.

The retrieval performance of three methods (the simple two-step, the candidate-keep two-step, and the binary file methods) is evaluated. SX flux waveforms are used for the evaluation. The number of waveforms is 10000. A waveform is divided into 450 segments. Each segment has 256 points. The simple two-step method asks the index to return 1000 segments. The candidate-keep two-step method uses 1000 as the initial candidate number, and 100 as the initial lower bound. For the binary file method, performance is measured under both of the hot and the cold states of the index. The hot state means that the whole of the index is on memory. The index is not on memory in the cold state. The retrieval time is measured by varying the number of segments of a query waveform. The result is shown in Fig. 1. This figure clarifies that the binary file method is the best. This may be caused by the number of consulting the multi-dimensional index. The binary file method consults the index only once.

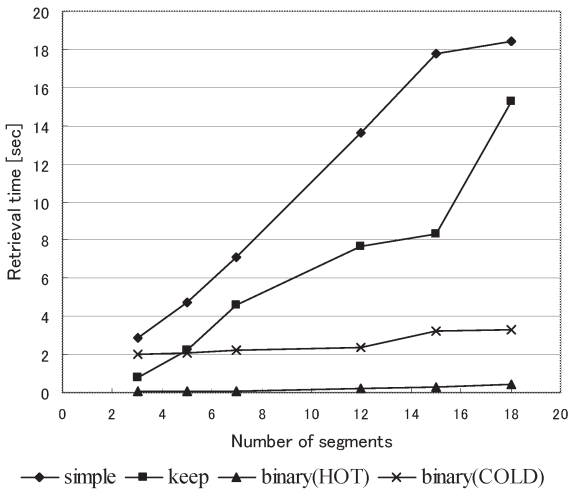

Fig. 1. Result of performance evaluation.

Reference

1) Hochin, T., Nakanishi H. and Kojima M.: On the Construction of Databases of Experiment Data, Ann. Rep. NIFS (2002) 167.